DOCUMENT RESUME

ED 422 385 TM 028 947

AUTHOR Althouse, Linda Akel; Ware, William B.; Ferron, John M.
TITLE Detecting Departures from Normality: A Monte Carlo Simulation of a New Omnibus Test Based on Moments.
PUB DATE 1998-04-00
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Monte Carlo Methods; *Sample Size; *Simulation; Statistical Distributions; Tables (Data)
IDENTIFIERS *Nonnormal Distributions; *Omnibus Tests; Power (Statistics)

ABSTRACT

The assumption of normality underlies much of the standard statistical methodology. Knowing how to determine whether a sample of measurements is from a normally distributed population is crucial both in the development of statistical theory and in practice. W. Ware and J. Ferron have developed a new test statistic, modeled after the K-squared test of R. D'Agostino and E. Pearson (1973), the g-squared test statistic. This statistic has been used to estimate critical values for sample sizes up to 100, but a more extensive derivation and validation of the critical values are required, and the power of g-squared against a wide range of alternative distributions requires study. Monte Carlo simulations were performed to investigate these areas. The main advantage of g-squared is its conceptual and computational simplicity. The power study shows that g-squared is sensitive to a wide range of alternative distributions, especially peaked distributions, having absolute power for many distributions with a large "n." G-squared could be valuable for testing univariate normality in statistical routines, but it does have some weaknesses. One of its main disadvantages is its low power with small sample sizes except for peaked distributions. While g- squared can tell you about a departure from normality, it can not tell if the departure is due to a single outlier. It is recommended that when testing for departures from normality, g-squared should be used as a supplemental quantitative measure of normality to the information obtained from histograms, box plots, stem and leaf diagrams, and normality plots. (Contains 7 figures, 12 tables, and 28 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 385

Detecting Departures from Normality: A Monte Carlo Simulation of a New Omnibus Test Based on Moments

Linda Akel Althouse
Columbia Assessment Services, Inc.

William B. Ware
University of North Carolina at Chapel Hill

John M. Ferron
University of South Florida

Paper presented at the Annual Meeting
of the
American Educational Research Association

San Diego, California
April 13-17, 1998

TM028947

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Linda Althouse

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Detecting Departures from Normality: A Monte Carlo Simulation of a New Omnibus Test Based on Moments

Linda Akel Althouse
Columbia Assessment Services, Inc.

William B. Ware
University of North Carolina at Chapel Hill

John M. Ferron
University of South Florida

Introduction

The assumption of normality underlies much of the standard statistical methodology employed for several reasons. First, many test statistics are assumed to be asymptotically normally distributed due to the applicability of large sample theorems such as the Central Limit Theorem. Second, the normal distribution is often assumed to be the appropriate mathematical model for underlying phenomena that the researcher may be investigating. That is, scores on many measures in the behavioral and social sciences are normally distributed so that the bell curve shape of the normal distribution provides a reasonable good fit to the frequency distributions of the scores. When using inferential statistics, this knowledge becomes useful as one can think of the distribution of the true magnitudes of a trait as being normally distributed in a population. Third, the normal curve provides a good approximation of other theoretical distributions that are more difficult to work with when determining probabilities. (D'Agostino, 1986; Glass & Hopkins, 1984; Shavelson, 1988).

With the assumption of normality yielding a rich set of mathematical consequences, it is no surprise that the normal distribution is the most widely used distribution in statistics. Therefore, knowing how to determine whether a sample of measurements is from a normally distributed population is crucial both in the development of statistical theory and in practice. As a result, much effort has been exerted in developing techniques solely for the purpose of detecting departures from normality. This effort began as early as the late 19th century with Pearson's (1895) work on moments, particularly the third and fourth moments which are commonly referred to as the skewness and kurtosis coefficients, respectively. However, while many tests currently exist, there is no gold standard among them as there is no one test which is both sensitive to a wide range of alternative distributions and

easy to compute. The variability in normality tests used is further evident by noting that even the major statistical packages such as SAS, SPSS, STATA, SYSTAT, and BMDP have implemented different normality tests (D'Agostino, Belanger, & D'Agostino, 1990; Hopkins & Weeks, 1990; Ware & Ferron, 1995).

Many would argue that Shapiro-Wilk (1965) W test is the most sensitive test to a wide range of alternative distributions. In fact, W was the first test for normality that was able to detect departures due to either skewness or kurtosis, or both. However, because of the complexity of this test, no statistical package has implemented W it its true form for sample sizes larger than 50. Rather large sample approximations (e.g., Royston, 1982) for W have been developed for use in statistical packages. Yet, it was the omnibus feature of the W test that motivated a new category of tests based on moments. For this new category of tests, the skewness and kurtosis coefficients are combined to provide an omnibus test for departures from normality. The most popular of these tests is the D'Agostino-Pearson (1973) K^2 test which is the sum of the squared normal approximations of the skewness and kurtosis coefficients. Being the sum of two normally distributed variables, K^2 is naturally distributed as χ^2 with 2 degrees of freedom. However, use of the χ^2 distribution requires that skewness and kurtosis be independent. Yet, studies have shown that they are related (Ware and Ferron, 1995; MacGillivray and Balanda, 1988).

In addition to vast number of tests available, Ware and Ferron (1995) also noted that the statistical packages report different estimates for skewness and kurtosis based on whether they are using Pearson or Fisher (1973) g estimates. Given these discrepancies, Ware and Ferron recommended an alternative test statistic, g^2, in which skewness and kurtosis coefficients from any of the leading statistical packages are combined to create an omnibus test for detecting departures from normality. Ware and Ferron developed this new test statistic, which was modeled after K^2, and estimated the critical values for sample sizes up to 100. However, a more extensive derivation and validation of the critical values is needed. In addition, the power of g^2 against a wide range of alternative distributions is currently unknown and must be determined. Both of these serve as the purpose of this study.

Motivation for Testing for Normality

In introductory statistics courses, one usually learns to test for normality by examining the distribution of the variables in question, comparing the mean, median, and mode of the distribution, and by examining outlier values. The importance of normality testing is stressed, yet formal tests are rarely introduced (Hopkins & Weeks,

1990).

Visually inspecting the data through the use of histograms, stem and leaf plots, and descriptive information is useful, yet it can only tell the researcher if the distribution is "close" to normal. Formal tests of the null hypothesis that the distribution is normal would provide one with a more precise indication of normality. Hopkins and Weeks (1990) noted that both descriptive and inferential measures of non-normality should be routine parts of reporting research. They noted a large number of robustness studies in the 1950s and 1960s that were designed to assess the consequences of non-normality. Glass, Peckham, and Sanders (1972) found that even though the assumption of normality is made in the derivation and use of parametric tests such as ANOVA, non-normality did not appear to have any serious consequences on the accuracy of the significance levels and the inferences made about the population mean. That is, it is frequently stated that ANOVA is robust to the assumption of normality (Tabachnick & Fiddell, 1996). However, Bradley (1978) questioned the generalization that inferential tests, such as ANOVA and the t-test, are robust to the violation of assumptions of normality since these generalizations sometimes neglect to mention qualifying conditions.

Hopkins and Weeks (1990) noted that because robustness is sometimes assumed, an unfortunate side effect resulted: normality is no longer considered or reported in research studies. Yet, there are still many statistical techniques in which the assumption of normality is crucial as some procedures are not robust to violations of the normality assumption (Hopkins & Weeks, 1990). For example, Breckler (1990) reviewed 72 articles that used structural equation modeling and less than 10% of the articles considered the crucial assumption of normality. In fact, with the increased use of structural equation modeling over the past decade, the issue of normality has become even more important as this statistical routine is not robust to the normality assumption (Bollen, 1989; Chou & Bentler, 1995; Hayduk, 1987; West, Finch, & Curran, 1995).

When using structural equation modeling, potential problems occur when nonnormal data are encountered when estimation techniques such as maximum likelihood or generalized least squares are used, as these estimation routines assume an underlying assumption of normality (West, Finch, & Curran, 1995). Therefore, statistical routines relying on these types of estimates will be affected by nonnormal data. In fact, when using structural equation modeling, it is important to test for both univariate and multivariate normality. Bollen (1989) cautioned that with nonnormal data, the chi-square estimate for assessing the fit of a structural equation model should be used with caution. This is particularly true for leptokurtic distributions that result in the rejection of too many true

models. Bollen also noted that for skewed distributions, the chi-squared estimates were also high. However, it is not clear whether these high values are due to skewness or to the naturally occurring kurtosis that tends to accompany skewness. Although estimation-based remedies do exist, it is important to know when your data are nonnormal so that the appropriate estimator can be used.

In addition to structural equation modeling, there are more common situations in which normality is required (e.g., regression, tests of variances, and meta-analysis. For example, in regression, it is necessary to assume that the residuals are normally distributed (Pedhazur, 1982). Tests of variance are also not robust to the violation of the normality assumption (Box, 1953). In meta-analysis, the homogeneity of effect sizes is crucial. That is, the estimates of effect sizes from a series of studies should be equal. However, the interpretation of the effect sizes depends on the assumption that the distribution follows a normal distribution (Glass, McGaw, & Smith, 1981; Hedges & Olkins, 1985; Wolf, 1986). For example, Greenhouse and Iyengar (1994) noted that for random effect models, when the distribution of effect sizes is skewed, the mean effect size might be positive while more than half of all the effect sizes are negative.

Given the information above on the importance of testing for the assumption of normality, it is important to have tests for normality. In fact, Tabachnick and Fiddell (1996) noted that normality is an important issue in data screening regardless of the inferential test being used. Currently, not all statistical packages have implemented procedures to test for normality, and, as mentioned earlier, the packages also differ in which test(s) they have implemented. Also, many of the tests for normality are not easy to implement, discouraging the average researcher from using them. The ideal test for normality would be easily computable from the output of any of the major statistical packages. Ware and Ferron's (1995) g^2 test statistic has these desirable properties. However, before it can be used an extensive derivation of the critical values and an analysis of its power must be conducted.

Statement of Purpose

This study was designed with two purposes. The first purpose was to extend the work of Ware and Ferron's (1995) empirical derivation of the critical values with a more extensive computer simulation. The second purpose was to determine the power of g^2 and compare it against the following test statistics: K^2, the standardized third moment test ($\sqrt{b_1}$), the standardized fourth moment test (b_2), and a large sample approximation of the W test (Royston, 1992). The first three competing tests were chosen as they are closely related to g^2. The test statistic

K^2 was the basis for developing g^2. The skewness and kurtosis tests are historically used in power studies for comparisons, particularly if the test statistic is derived as a combination of the two measures. The Shapiro-Wilk W approximation was chosen because it is currently regarded as being the most powerful test for a number of alternative distributions.

Defining g^2

Ware and Ferron (1995) defined g^2 as

$$[Z(g_1)]^2 + [Z(g_2)]^2 \qquad \text{(Formula 1)}$$

$$Z(g_1) = g_1 / SE(g_1) \qquad \text{and} \qquad Z(g_2) = g_2 / SE(g_2).$$

While Ware and Ferron chose to express this statistic in terms of the Fisher estimates, g^2 can easily be computed with either the Pearson or Fisher estimates as long as the standard error of the estimates are available. Fisher estimates were used instead of Pearson estimates as most computer packages use the Fisher estimates when reporting skewness and kurtosis. For those using packages reporting Pearson estimates rather than Fisher, it can be shown that

$$g_1 / SE(g_1) = \sqrt{b_1}/SE(\sqrt{b_1}) \quad \text{and} \quad g_2 / SE(g_2) = (b_2 - E(b_2))/SE(b_2). \qquad \text{(Formula 2)}$$

Therefore for packages providing the Fisher estimates and standard errors, the g^2 value can be easily computed by using Formula 1. If the Pearson estimates are provided, then g^2 value can be computed applying Formula 2. Ware and Ferron (1995) provide examples of computing g^2 from five statistical packages. Once the value of g^2 is obtained, the next step is to determine if this value is significant indicating a departure from normality. To determine the significance of g^2, we generated the empirical distribution so that critical values for sample sizes up to n=5000.

Determining the Empirical Distribution of g^2

To obtain the critical values for g^2 a Monte Carlo simulation was conducted. Random samples were generated from the standard normal distribution using the SAS RANNOR function within PROC IML (SAS, 1995). Using 500,000 replications for sample sizes, n=10(1)100(25)500(50)1000(250)5000, g^2 was calculated. For each of the 143 sample sizes, the empirical critical values at the .10, .05, and .01 significance levels were estimated by using PROC UNIVARIATE to determine the point which was exceeded by 10%, 5%, and 1% of the g^2 values, respectively.

For example, consider a sample of size 10. To derive the estimated critical values at the above significance levels, a sample size of 10 was randomly generated from a normal population, and g^2 was calculated for this sample. This procedure was repeated for a total of 500,000 samples of size 10. Next, the 90th, 95th, and 99th percentiles for the 500,000 g^2 values were determined. These values were the estimated critical values for g^2 the .10, .05, and .01 significance levels, respectively.

Once the estimated critical values of g^2 were empirically derived for each n, these values were plotted as a function of sample size as can be seen in Figure 1.

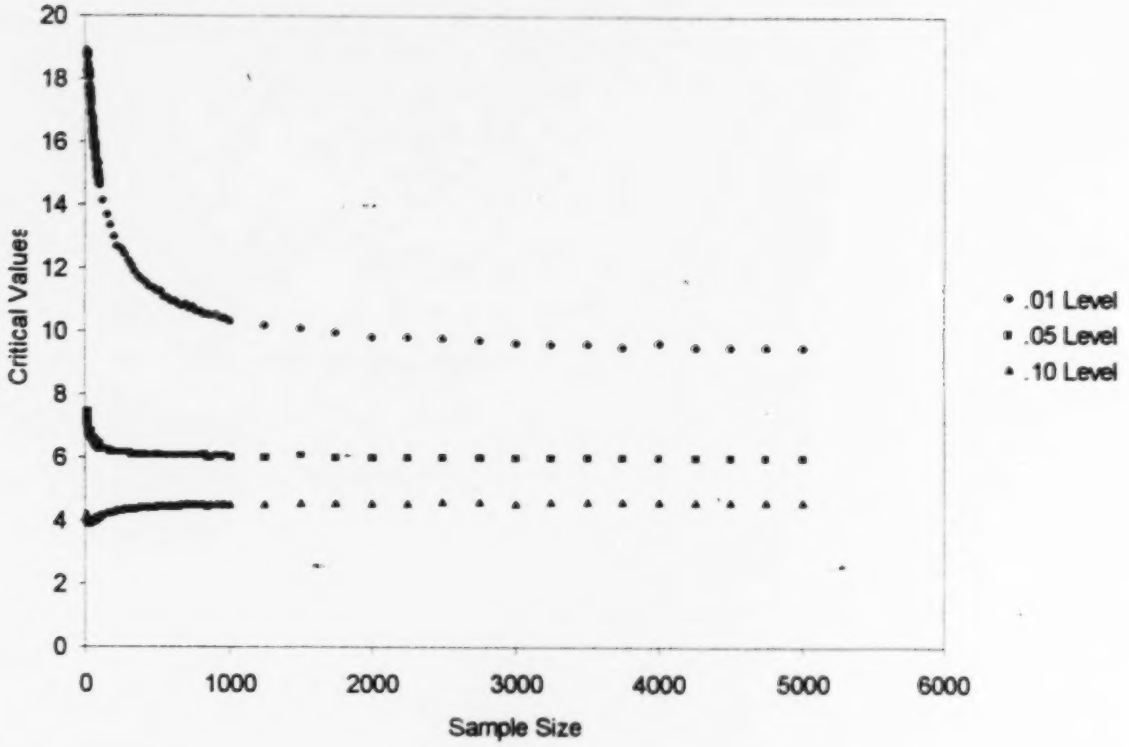

Figure 1: Plots of the Estimated Critical Values of g2 at the Three Levels of Significance as a Function of Sample Size

Examination of the above plots indicated that a nonlinear relationship existed between the sample sizes and the critical values. Linear, inverse, natural logarithm, cubic, quadratic, exponential, and gamma functions were used in several regression combinations to determine the model of best fit. These best fit models were then used to determine the empirical distribution of g^2. The resulting regression equations, based on the 143 critical value estimates at each significance level, are given below:

$$Y'_{.01} = -3.741696 + 1.2660111\ln(n) - 23.808164\left(\frac{1}{\ln(n)}\right) - 92.874186\left(\frac{1}{n}\right) - 15.247724\left(\frac{1}{\sqrt{n}}\right) + 94.06006\left(\frac{1}{\sqrt[3]{n}}\right)$$

$$Y'_{.05} = 14.093477 - 0.506116\ln(n) - 23.948084\left(\frac{1}{\ln(n)}\right) - 56.880456\left(\frac{1}{n}\right) + 80.242742 + \left(\frac{1}{\sqrt{n}}\right) - 35.834511\left(\frac{1}{\sqrt[3]{n}}\right) + 187.756861\left(\frac{1}{n^2}\right)$$

$$Y'_{.10} = 4.313216 - 0.041669\ln(n) - 0.623687\left(\frac{1}{\ln(n)}\right) - 0.431729\left(\frac{1}{\exp\left(n/200\right)}\right) - 0.005997\left(\frac{1}{\Gamma\left(n/200\right)}\right) + 60.681903\left(\frac{1}{\sqrt{n}}\right) - 153.163890\left(\frac{1}{\sqrt[3]{n}}\right)$$

These models accounted for 99.9%, 99.6%, and 99.7% of the variance in the three sets of estimated critical values. Using these regression equations, the estimated critical values of g^2 at the .10, .05, and .01 levels of significance were estimated and are provided in Table 1. The critical values for sample sizes not listed in this table can be obtained by substituting the sample size for n in the regression equations critical values above.

While it is expected that the critical values decrease as the sample size increase, deviations from this pattern were found in the set of critical values for g^2, particularly at the .10 level, necessitating further examination of these values. The additional investigation of the critical values suggested that the distribution of g^2 approached the χ^2 distribution with 2 degrees of freedom indicating that the distribution of g^2 was asymptotic to the this particular χ^2 distribution. Intuitively, we expect this relationship to occur since as n gets larger, the standardized skewness and kurtosis values used to compute g^2 become normally distributed making g^2 the sum of two normally distributed variables. To verify this relationship, separate Monte Carlo simulations were conducted for n=10, 100, 500, 1000, and 5000. In each of these simulations, 10,000 random samples were generated from the standard normal distribution and g^2 was calculated for each of the samples. The distribution of these g^2 values were plotted and compared to the distribution of 10,000 random χ^2 variates of the same sample size. As the sample size increased, the g2 distribution began to converge to the chi-squared distribution with 2 df. Therefore, for large sample sizes one can use the critical values for the χ^2 with 2 degrees of freedom.

Table 1: Listing of Critical Values for g^2 (alpha = .01, 05, & .10)

Sample Size	α = .01	α = .05	α = .10
10	18.3835	7.4592	4.1816
11	18.6186	7.3548	4.1312
12	18.7667	7.2741	4.0923
13	18.8534	7.2094	4.0621
14	18.8959	7.1559	4.0385
15	18.9062	7.1104	4.0200
16	18.8929	7.0709	4.0054
17	18.8619	7.0359	3.9939
18	18.8177	7.0046	3.9849
19	18.7638	6.9760	3.9779
20	18.7026	6.9498	3.9724
30	17.9562	6.7612	3.9627
40	17.2452	6.6392	3.9812
50	16.6432	6.5504	4.0053
60	16.1370	6.4822	4.0298
70	15.7069	6.4280	4.0535
80	15.3367	6.3839	4.0760
90	15.0141	6.3472	4.0972
100	14.7299	6.3164	4.1171
125	14.1454	6.2570	4.1621
150	13.6890	6.2147	4.2011
175	13.3202	6.1830	4.2352
200	13.0141	6.1586	4.2650
225	12.7549	6.1394	4.2911
250	12.5319	6.1238	4.3142
275	12.3375	6.1110	4.3346
300	12.1661	6.1004	4.3526
325	12.0137	6.0914	4.3686
350	11.8769	6.0838	4.3828
375	11.7535	6.0773	4.3955
400	11.6413	6.0717	4.4067
425	11.5389	6.0668	4.4168
450	11.4449	6.0625	4.4258
475	11.3584	6.0587	4.4339
500	11.2783	6.0553	4.4411
550	11.1349	6.0497	4.4536
600	11.0099	6.0451	4.4639
650	10.9000	6.0415	4.4726
700	10.8024	6.0384	4.4799
750	10.7151	6.0359	4.4862
800	10.6366	6.0337	4.4916
850	10.5656	6.0319	4.4965
900	10.5011	6.0304	4.5008
950	10.4421	6.0290	4.5047
1000	10.3880	6.0278	4.5082
1500	10.0228	6.0211	4.5325
2000	9.8271	6.0175	4.5479
2500	9.7090	6.0146	4.5595
3000	9.6333	6.0116	4.5689
3500	9.5830	6.0086	4.5768
4000	9.5493	6.0054	4.5836
4500	9.5268	6.0022	4.5896
5000	9.5123	5.9989	4.5949
χ^2_2	9.2104	5.9915	4.6052

Validation of g^2 and K^2

Even though the variance accounted for in the three sets of observed critical values was high, the set of critical values for g^2 generated by the regression equations was further validated by conducting a smaller Monte Carlo simulation. This validation provided an estimate of the Type I error rate for g^2. In addition to validating the critical values of g^2, the D'Agostino-Pearson K^2 test was also validated. K^2 was validated in order to help determine whether the relationship between skewness and kurtosis had an adverse impact on the accuracy of the critical values for the K^2 test. While there have been numerous power studies conducted on K^2, no studies were found that provided an empirical investigation of the violation of independence or an examination of its Type I error rate. In addition, since g^2 was developed as a modification of K^2, it seemed appropriate to validate K^2 along with the new test statistic.

To perform the validation, 10,000 random samples were generated from a normal population for the following 45 sample sizes, n=10(1)15(5)25(1)30(10)100(50)650(25)1000. In addition to wanting a representative range of sample sizes from 10 to 1000, these sample sizes were selected to include sample sizes where the potential problem of the increasing critical values occurred. At each significance level and for each sample size, the proportion of times that the calculated g^2 value exceeded the critical values obtained from the regression was determined. In addition, the proportion of times that the calculated K^2 exceeded the critical values based on the chi-squared distribution was recorded. The proportions for each sample size for both g^2 and K^2 are provided in Table 2.

Table 2: Validation of g^2 and K^2 - Proportion of Times Departures from Normality Detected

n	.10		.05		.01	
	g^2	K^2	g^2	K^2	g^2	K^2
10	.098	.093	.049	.057	.009	.021
11	.100	.089	.050	.057	.010	.022
12	.101	.093	.054	.061	.010	.025
13	.097	.090	.050	.057	.011	.024
14	.103	.094	.051	.057	.010	.023
15	.097	.091	.048	.055	.010	.021
20	.098	.093	.049	.056	.011	.021
25	.096	.092	.047	.053	.010	.020
26	.097	.094	.048	.055	.009	.020
27	.101	.097	.051	.060	.011	.020
28	.100	.095	.051	.059	.011	.021
29	.098	.096	.050	.057	.010	.019
30	.098	.094	.048	.056	.009	.019
40	.095	.094	.048	.055	.008	.018
50	.103	.101	.052	.059	.008	.017
60	.102	.101	.049	.057	.009	.018
70	.101	.100	.049	.057	.011	.018
80	.098	.101	.048	.055	.010	.019
90	.104	.102	.052	.058	.011	.018
100	.097	.097	.047	.054	.011	.018
150	.103	.102	.051	.053	.011	.016
200	.101	.099	.050	.053	.010	.016
250	.097	.095	.050	.053	.011	.016
300	.098	.098	.051	.053	.011	.016
350	.101	.100	.050	.052	.009	.013
400	.099	.097	.051	.051	.010	.012
450	.095	.097	.050	.051	.010	.013
500	.099	.099	.049	.052	.011	.014
550	.104	.103	.052	.052	.012	.015
600	.099	.095	.048	.047	.010	.012
650	.102	.102	.051	.053	.010	.013
675	.099	.099	.050	.050	.009	.011
700	.104	.103	.052	.054	.011	.013
725	.102	.100	.045	.048	.009	.010
750	.097	.098	.048	.050	.010	.012
775	.104	.103	.053	.053	.010	.012
800	.099	.100	.050	.049	.010	.012
825	.100	.100	.050	.051	.010	.012
850	.101	.102	.050	.053	.010	.013
875	.106	.106	.055	.057	.010	.012
900	.102	.100	.050	.053	.010	.013
925	.099	.099	.049	.052	.010	.011
950	.102	.100	.050	.054	.009	.012
975	.093	.094	.044	.048	.008	.010
1000	.100	.098	.047	.049	.011	.013
Mean	.09972	.09766	.04971	.05384	.00993	.01604
sd	.00284	.00401	.00202	.00332	.00094	.00418
z-score	-.65516	-3.9142	-.9593	7.7647	-.50669	9.6960

The proportions of the g^2 statistics, averaged over the 45 sample sizes, falling above the critical values at the .10, .05, and .01 significance levels were .0997 (sd=.0028), .0497 (sd=.0020), and .0099 (sd=.0001), respectively . Converting these proportions to z-scores, we found that these values were not significantly different from the expected proportions .10, .05, and .01 (z=-.6552, -.9593, and -.5067, respectively) indicating that the Type I error rates were as expected for each level of significance.

Similarly, the mean proportion of K^2 statistics falling above the critical values at the .10, .05., and .01 significance levels were .0977 (sd=.0040), .0538 (sd=.0033), .0160 (sd=.0042). As with the g^2 test, these proportions were converted to z-scores. Each of the mean proportions was found to differ significantly from the expected proportions .10, .05, and .10 (z= -3.9142, 7.7647, and 9.9690, respectively). Specifically, at the .10 level, the Type I error rate was significantly lower than the expected value of .10. At the .05 and .01 levels, K^2 had an inflated Type I error rate indicating that it may detect departures from normality when a sample is actually normally distributed.

Since the proportions validated for g^2 were based on the mean of all the sample sizes, it is possible that significant differences at specific sample sizes were present but were undetected with the balancing of underestimates and overestimates. In order to determine if the 45 sample sizes and the frequencies above and below the critical values at each of these sample sizes were statistically independent of each other, a chi-squared test was conducted. That is, the frequencies above and below the critical values were compared to the expected frequencies, stratifying by sample size. For each of the significance levels (.10, .05, .01), the resulting χ^2 statistic with 44 degrees of freedom was not statistically significant at the .05 level ($p<.658$, $p<.725$, $p<.658$, respectively) indicating statistical independence. This follow up χ^2 test was not conducted for K^2 since a significant difference was found for the aggregate proportion values.

The above efforts demonstrated that the g^2 test statistic did validate and had the expected Type I error rate. However, given that K^2 did not validate, we were prompted to run a smaller validation study to see what patterns emerged for the other three test statistics. As with K^2, we found no studies reporting the Type I error rates for these tests. In order to directly compare the validation information of g^2 and K^2 with the other three test statistics, g^2 and K^2 were included in this second validation step.

Analysis of Type I Error Rates

In this simulation, 10,000 samples were generated from the standard normal population for the following twelve sample sizes, n=10, 25, 50, 75, 100, 150, 200, 300, 400, 400, 750, and 1000. The five test statistics were calculated for each sample. At each significance level and for each sample size, the proportion of times that each statistics exceeded their defined critical values was determined. The proportions for each sample size for each test statistic are provided in Table 3.

The proportions for g^2 were similar to the theoretical proportions for each significance level, once again, validating the use of the regression equations and indicating no problems with Type I error rates. As seen earlier, for K^2, there was a slight, yet consistent, inflation of Type I error rates at the .01 and .05 levels of significance, particularly for small sample sizes. The Type I error rates for $\sqrt{b_1}$ and b_2 were fairly close to their expected values. However, most troublesome were the Type I error rates for W approximation which were largely inflated at all significance levels. As the sample size increased, these Type I error rates were more exaggerated. Overall, W was the most liberal of the six tests.

Once the Type I error rates were examined, the next step was to evaluate the powers of g^2 to determine how well it would detect departures from normality and to see how its power compared to those of the competing test statistics. However, given the inflated Type I error rates for K^2 and W, we entered this next phase suspecting that the power of K^2 and W would be somewhat inflated, making them appear more powerful.

Table 3: Comparison of Type I Error Rates for Five Competing Test Statistics

Level of Significance	n	g^2	K^2	$\sqrt{b_1}$	b_2	W
α=.10	10	.105	.096	.102	.090	.098
	25	.103	.099	.102	.100	.119
	50	.097	.098	.101	.102	.124
	75	.103	.103	.104	.102	.125
	100	.100	.097	.099	.104	.129
	150	.097	.096	.097	.105	.132
	200	.102	.102	.101	.102	.141
	250	.099	.097	.099	.099	.144
	300	.098	.099	.099	.098	.140
	400	.099	.096	.099	.103	.153
	500	.099	.102	.103	.109	.157
	750	.101	.102	.099	.105	.160
	1000	.099	.100	.100	.097	.170
Mean Type I Error Rate at α=.10		.1002	.0990	.1004	.1012	.1378
α=.05	10	.051	.061	.054	.044	.053
	25	.050	.060	.052	.053	.063
	50	.049	.057	.051	.054	.066
	75	.052	.059	.052	.056	.073
	100	.049	.054	.047	.055	.072
	150	.049	.054	.047	.055	.074
	200	.053	.055	.051	.054	.084
	250	.049	.051	.047	.051	.084
	300	.048	.052	.048	.050	.087
	400	.050	.051	.049	.052	.093
	500	.051	.053	.049	.056	.097
	750	.051	.053	.051	.054	.104
	1000	.050	.051	.053	.051	.113
Mean Type I Error Rate at α=.05		.0502	.0547	.0501	.0527	.0818
α=.01	10	.012	.025	.013	.007	.013
	25	.011	.022	.011	.011	.016
	50	.009	.018	.009	.011	.014
	75	.011	.019	.011	.014	.021
	100	.012	.019	.010	.015	.020
	150	.009	.016	.010	.012	.022
	200	.010	.014	.010	.012	.025
	250	.011	.014	.010	.012	.025
	300	.010	.013	.008	.013	.027
	400	.010	.014	.010	.012	.034
	500	.009	.014	.011	.012	.032
	750	.013	.014	.011	.012	.040
	1000	.011	.013	.010	.012	.045
Mean Type I Error Rate at α=.01		.1016	.0165	.0103	.0119	.0257

The Power Study

Choosing the Alternative Distributions

The first step in any power study is to determine the alternative distributions against which the desired test statistic will be evaluated. Lists of alternative distributions with their skewness and kurtosis population values have been provided by many authors who have conducted previous power studies (Albajar, Moreno, & Martin, 1992; Pearson, D'Agostino, & Bowman, 1977; Shapiro, Wilk, & Chen, 1968; Saniga & Miles, 1979; Stephens, 1974). These lists were used in determining the subset of alternative distributions used for this study. The main intent was to select a varied set of distributions so that the power of g^2 could be examined against many different types of departures from normality. The final set of alternative distributions chosen included symmetric and skewed distributions with low (flat) and high (peaked) kurtosis values, as well as, distributions close to the normal distribution. With the exception of near normal distributions, which considered both continuous and discrete distributions, all the alternative distributions considered were continuous. These choices were reflective of the types of distributions utilized in previous power studies.

Two distributions were selected for each of the following six categories of distributions: near normal discrete, near normal continuous, symmetric/flat, symmetric/peaked, skewed/flat, and skewed/peaked. The resulting twelve distributions, along with their population skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) values, are presented in Table 4. The skewness and kurtosis values shown in the table indicate the degree of departure from normality, where $\sqrt{\beta_1}=0$ and $\beta_2 = 3$. SAS 6.11 was used for the power study and the coding of the alternative distributions.

Table 4. Categorization of the Alternative Distributions Used in the Power Study

Distribution Category	Alternative Distributions (parameters)	$\sqrt{\beta_1}$	β_2
Near Normal Discrete	Binomial (20, .5)	0	2.90
	Poisson (10)	.32	3.10
Near Normal Continuous	Tukey (1, 5)	0	2.90
	Johnson S Bounded (1, 2)	.28	2.77
Symmetric/flat	Johnson S Bounded (0, .5)	0	1.63
	Tukey (1, 1.5)	0	1.75
Symmetric/peaked	Johnson S Unbounded (0, 2)	0	4.71
	Johnson S Unbounded (0, .9)	0	82.08
Skewed/flat	Johnson S Bounded (.533, .5)	.65	2.13
	Beta (2, 1)	-.57	2.40
Skewed/peaked	Johnson S Unbounded (1, 1)	-5.30	93.40
	Lognormal (0, 1, 0)	6.18	113.94

Computing the Power of g^2

To determine the power of g^2 and compare it to the powers of $\sqrt{\beta_1}$, β_2, K^2, and W, another Monte Carlo empirical sampling study was conducted. The 12 alternative distributions listed in Table 4 were considered for each of the following sample sizes (n=10, 25, 50, 75, 100, 150, 200, 250, 300, 400, 500, 750, 1000) at the .10, .05, and .01 significance levels. For each alternative distribution, 10,000 iterations of samples were generated for each sample size. The test statistics were calculated for each of the 10,000 samples. The number of times a test statistic detected a deviation from normality was tracked and converted to a percentage value, providing the estimated power of the test.

Factorial Analysis of the Power Results

In order to determine the various effects, particularly interaction effects, on the power of the test statistics, the obtained power results were evaluated using a four-way (3 x 5 x 6 x 13) factorial analysis (level of significance, test statistic, distribution category, and sample size). The independent variables used and their possible values are defined in Table 5. The dependent variable was the power of the test statistic.

Table 5: Independent Variables used in the Four-Way Factorial Analysis of the Power Results

Independent Variables	Possible Values
Level of Significance	.10, .05, .01
Type of Test Statistic	g^2, K^2, $\sqrt{b_1}$, b2, W
Category of Distribution	Near Normal Discrete, Near Normal Continuous, Symmetric/Flat, Symmetric/Peaked, Skewed/Flat, Skewed/Peaked
Sample Size	10, 25, 50, 75, 100, 150, 200, 250, 300, 400, 500, 750, 1000

The summary of the results of the ANOVA is presented in Table 6. The highest order significant interaction effect was a 3-way interaction which indicated an there was an effect on power due to the interaction of the type of test statistic, the distribution category, and sample size. The significant three-way interaction implied that the interaction of two of the independent variables depended on the specific level of the third independent variable. To examine this effect, the interaction between two of the independent variables at each level of the third independent variable was analyzed graphically. More specifically, the interaction effect between the test statistic and sample size was graphed for each of the six types of distribution categories. For each distribution category, the two power values were averaged across the three significance levels yielding a mean power estimate for each test statistic for each of the 13 sample sizes.

Table 6: Four-Way ANOVA Summary Table with Effect Sizes for Investigating the Relationships between Significance Levels, Distribution Category, Sample Size, Test Statistic, and Power

Source of Variations	Sum of Squares	df	Mean Squares	F	η^2
Main Effects					
Alpha	56469.19	2	28234.60	66.42*	.01
Dist	1434601.66	5	286920.33	674.93*	.35
Size	828641.33	12	69053.44	162.44*	.21
Test	315051.03	4	78762.76	185.28*	.08
2-Way Interactions					
Alpha Dist	7093.91	10	709.39	1.67	.00
Alpha Size	12080.02	24	503.33	1.18	.00
Alpha Test	3956.49	8	494.56	1.16	.00
Dist Size	165297.90	60	2754.97	6.48*	.04
Dist Test	506180.15	20	25309.01	59.54*	.12
Size Test	42833.43	48	892.36	2.10*	.01
3-Way Interactions					
Alpha Dist Size	21038.01	120	447.94	.41	.00
Alpha Dist Test	6563.19	40	175.32	.39	.00
Alpha Size Test	9228.02	96	164.08	.23	.00
Dist Size Test	185349.07	240	96.13	1.82*	.04
4-Way Interactions					
Alpha Dist Size Test	37151.14	480	77.40	7.31	.01
Error	497382.86	1170	425		
Total	4128917.41	2339			

* $p < .01$

Results

Near Normal Discrete Distributions

The near normal discrete distributions consisted of the binomial (20, .5) distribution and the Poisson (10) distribution. Each of these distributions had a low to zero skewness value and a kurtosis value close to three. The mean power of each test statistic for each sample size is presented in Table 7 and displayed graphically in Figure 2.

Table 7: Near Normal Discrete Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	5.28	5.52	4.43	5.82	9.02
25	5.82	6.42	5.60	6.55	15.50
50	7.28	8.48	6.60	8.27	28.83
75	8.27	10.62	6.72	9.48	45.60
100	9.65	12.75	7.07	11.03	62.92
150	12.60	17.17	7.28	14.15	86.32
200	15.38	20.95	7.75	17.40	95.93
250	18.55	25.20	7.77	20.62	99.12
300	22.18	28.98	8.33	24.42	99.92
400	27.90	35.23	8.47	30.53	100
500	32.95	40.07	8.92	35.85	100
750	43.10	47.28	10.52	45.33	100
1000	48.60	50.40	11.40	50.52	100

The test statistic g^2 did not perform as well as W for near normal discrete distributions. However, the power of g^2 was greater than the power of the kurtosis test which exhibited the lowest power for this set. In addition, g^2 had similar power to K^2 and $\sqrt{b_1}$. The Royston approximation for W was clearly the most powerful test.

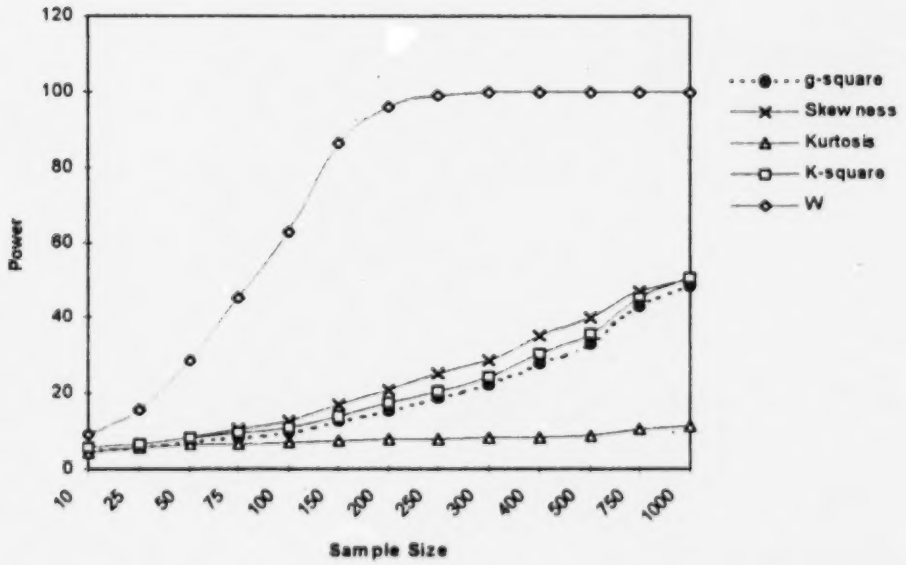

Figure 2: Mean Powers for the Near Normal Discrete Distributions as a Function of Sample Size and Test Statistic

Near Normal Continuous Distributions

The near normal continuous distributions consisted of the Tukey (1,5) distribution and the Johnson S Bounded (1, 2) distribution. As with the above discrete functions, each of these distributions had a low to zero skewness value and a kurtosis value close to three. The mean power of each test statistic for each sample size is given in Table 8 and presented graphically in Figure 3.

As with the discrete near normal distributions, g^2 did not perform as well as W for this set. In fact, g^2 had lower power for all sample sizes. The test statistics K^2, $\sqrt{b_1}$, and b_2 had equally low power as g^2. The Royston approximation for W once again demonstrates the highest power, particularly for larger sample sizes.

Table 8: Near Normal Continuous Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	5.83	5.72	4.83	6.10	6.25
25	4.15	4.40	4.20	4.35	8.70
50	3.63	5.08	4.58	4.75	14.67
75	4.10	6.68	4.72	5.67	23.55
100	5.08	8.40	5.20	7.05	34.02
150	7.87	12.33	6.02	1.48	53.83
200	11.37	16.62	6.80	14.68	68.75
250	15.33	20.43	7.62	19.18	78.18
300	19.25	24.05	8.27	23.63	84.07
400	26.50	30.87	9.93	31.85	91.55
500	33.20	36.72	11.42	38.42	96.05
750	44.43	44.97	15.05	47.20	99.45
1000.	48.92	48.47	19.35	49.97	99.97

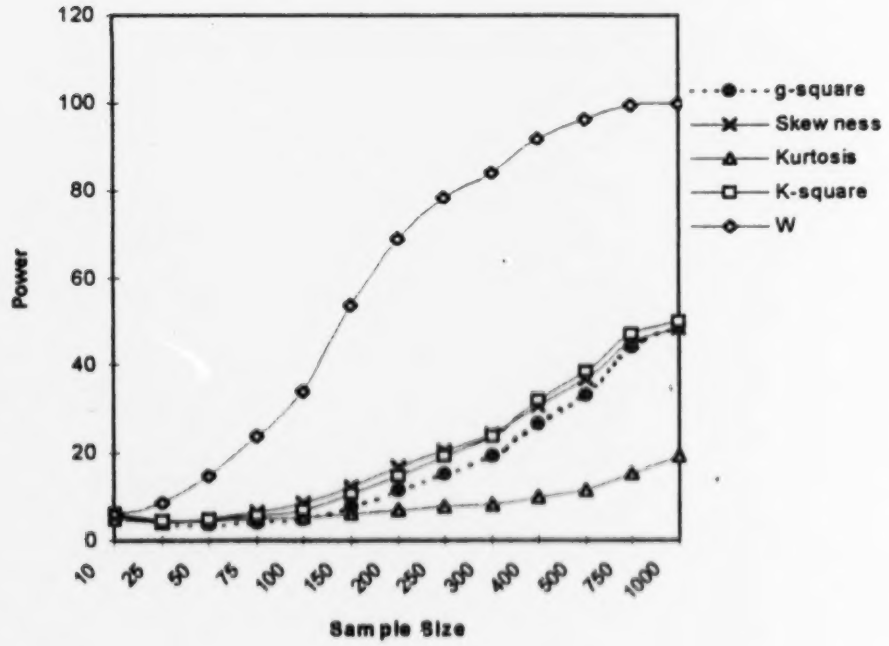

Figure 3: Mean Powers for the Near Normal Continuous Distributions as a Function of Sample Size and Test Statistic

Symmetric/Flat Distributions

The set of distributions included the Johnson S Bounded (0, .5) and the Tukey (1, 1.5) distributions. Both of these distributions have skewness values equal to zero and less than three. The mean powers for this set of distribution are provided in Table 9 and Figure 4.

The test statistic g^2 did not appear as powerful as either W or K^2 for this set until the sample size reached over 250. For large sample sizes, g^2 was as good as the leading tests. The test statistic b_2 had the highest power with W and K^2 equally close behind. The skewness test had the weakest power.

While g^2 is weak for this set, its power is higher for this group than for the two near normal sets. One notable difference in this graph and the graphs for the near normal distributions is the rate at which the power increases. The slopes of the functions for the symmetric/flat distributions, with the exception of the skewness test, were steeper than the

slopes of the normal distributions, particularly for sample sizes less than 75. That is, the test statistics gained power at a much quicker rate for symmetric/flat distributions than near normal distribution, making g^2 an option for large sample sizes.

Table 9: Symmetric/Flat Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	1.67	2.67	10.57	4.57	13.32
25	.63	.90	54.52	39.45	47.37
50	18.68	.57	92.97	88.77	88.42
75	43.00	.48	99.18	98.58	98.57
100	60.95	.35	99.95	99.87	99.88
150	69.57	.37	100	100	100
200	93.18	.32	100	100	100
250	99.77	.30	100	100	100
300	100	.30	100	100	100
400	100	.33	100	100	100
500	100	.32	100	100	100
750	100	.27	100	100	100
1000	100	.28	100	100	100

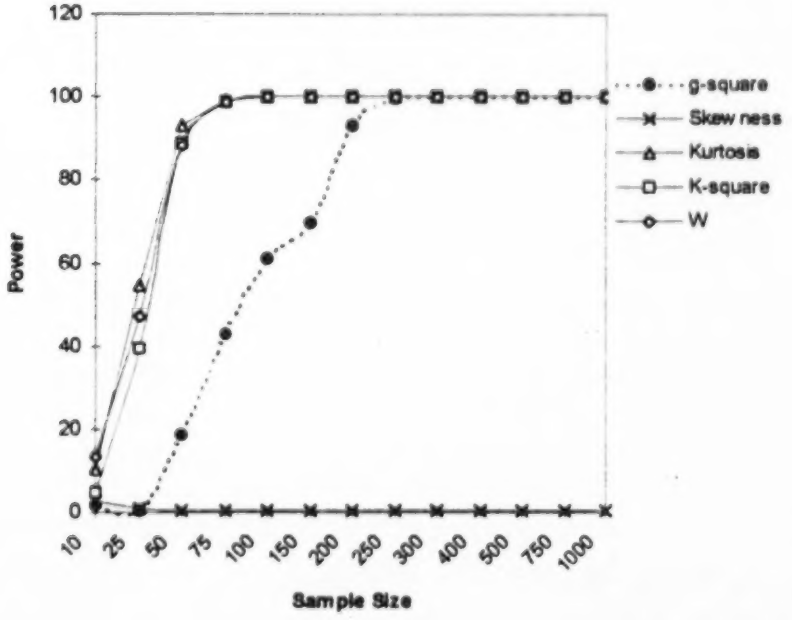

Figure 4: Mean Powers for the Symmetric/Flat Distributions as a Function of Sample Size and Test Statistic

Symmetric/Peaked Distributions

The symmetric/peaked set consisted of Johnson S Unbounded (0, 2) and the Johnson S Unbounded (0, .9) distributions. Both of these distributions had kurtosis values over three. The mean powers for this set of distributions are presented in Table 10 and Figure 5.

The test statistic g^2 was as powerful or more powerful than all other test statistics for all sample sizes. The power of g^2 was very similar to K^2, b_2, and W. As with the symmetric/flat distributions, $\sqrt{b_1}$ provided the weakest power. Also, the mean powers for g^2, K^2, b_2, and $\sqrt{b_1}$ were notably higher from what they were in the previous sets of distributions. W had more power for smaller sample sizes and less power for larger sample sizes than it did with near normal distributions. However, in comparison to the symmetric/flat distributions, the mean power of W was consistently lower for this set of distribution.

Table 10: Symmetric/Peaked Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	19.80	19.33	15.85	20.95	17.43
25	38.83	32.33	33.52	37.93	36.50
50	55.32	41.13	50.72	53.05	53.87
75	64.32	45.80	60.33	61.50	62.48
100	69.32	48.75	66.12	66.50	67.67
150	75.72	52.30	73.03	72.92	73.97
200	80.35	53.95	78.05	77.28	78.83
250	84.60	55.87	82.75	81.58	83.43
300	87.22	56.62	85.75	84.30	86.33
400	91.77	58.47	91.18	89.40	91.42
500	94.85	59.13	94.58	92.97	94.88
750	98.62	61.78	98.55	97.82	98.77
1000	99.67	63.12	99.65	99.43	99.75

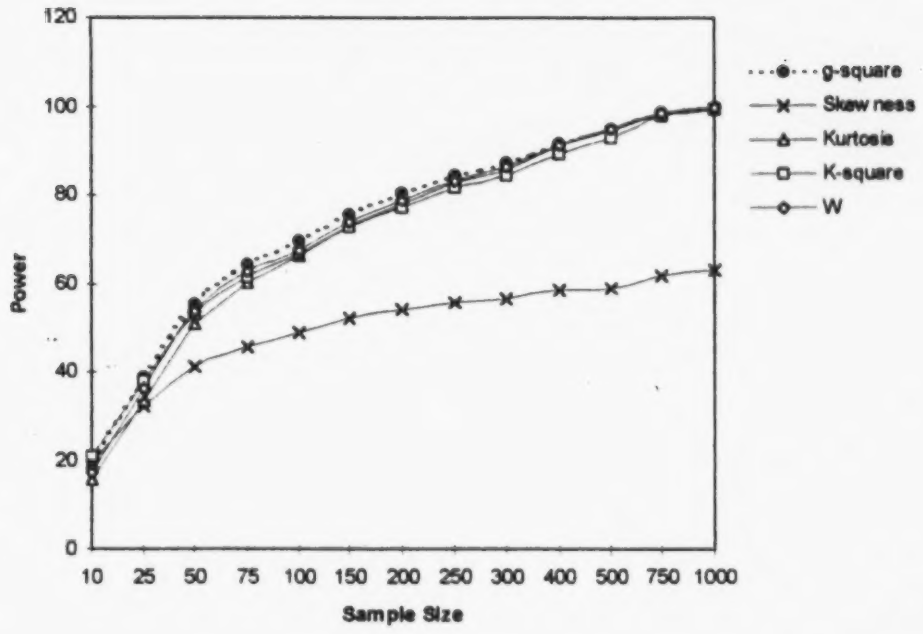

Figure 5: Mean Powers for the Symmetric/Peaked Distributions as a Function of Sample Size and Test Statistic

Skewed/Flat Distributions

This set of distributions included the Johnson S Bounded (.533, .5) and the Beta (2, 1) distributions which are positively and negatively skewed, respectively. In addition, both these distributions have a kurtosis value less than three. The mean powers for this set are presented in Table 11 and Figure 6.

As with the earlier flat shaped distributions, W is clearly the most powerful, particularly for sample sizes up to 200. For larger sample sizes, g^2, K^2, $\sqrt{b_1}$, and W had equivalent power. The kurtosis test statistic b_2, was the weakest test, especially for sample sizes less than 500. The powers of g^2 were higher for this set than the reported powers for the near normal and symmetric/flat distributions. However, they were lower than the powers found for the symmetric/peaked distributions. For small sample sizes, g^2 had the weakest power.

Table 11: Skewed/Flat Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	7.22	10.48	9.25	9.58	21.98
25	10.13	18.53	19.75	21.88	61.52
50	30.78	36.50	33.17	53.75	91.02
75	52.28	54.75	43.32	79.25	98.55
100	62.95	69.12	52.22	89.02	99.85
150	83.02	88.05	65.17	98.77	100
200	94.50	96.17	74.93	99.95	100
250	99.50	98.87	81.70	100	100
300	99.98	99.73	86.80	100	100
400	100	100	92.87	100	100
500	100	100	96.18	100	100
750	100	100	99.28	100	100
1000	100	100	99.88	100	100

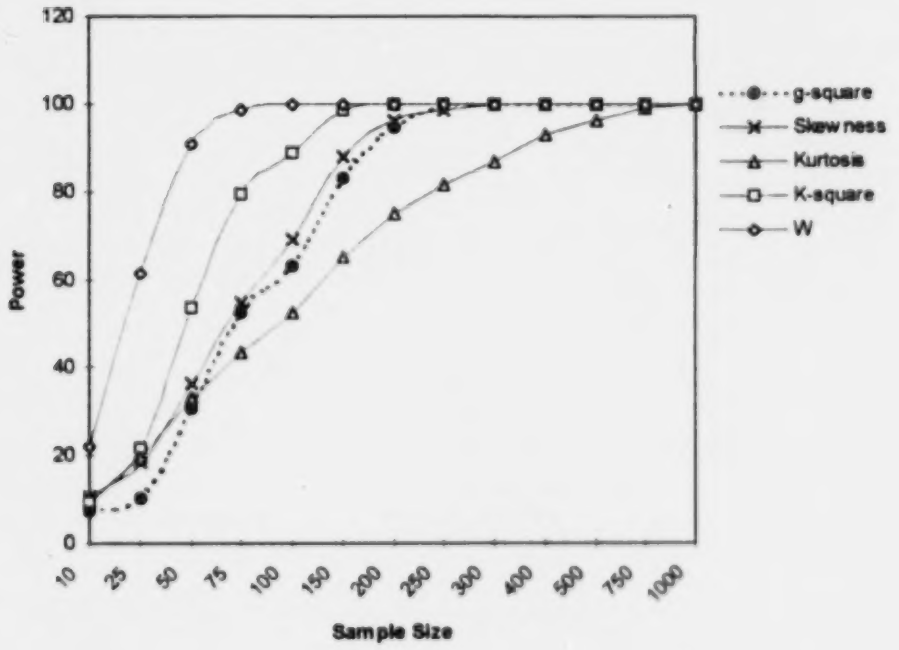

Figure 6: Mean Powers for the Skewed/Flat Distributions as a Function of Sample Size and Test Statistic

Skewed/Peaked Distributions

This category included the Johnson S Unbounded (1, 1) and the lognormal (0, 1, 0) distributions which are positively and negatively skewed, respectively. In addition, both these distributions have a kurtosis value larger than 3. The mean powers for this set are presented in Table 12 and Figure 7.

Table 12: Skewed/Peaked Distributions - Estimated Mean Power for Test Statistics by Sample Size

Sample Size	g^2	$\sqrt{b_1}$	b_2	K^2	W
10	38.57	43.83	30.35	42.10	47.83
25	77.60	84.07	62.10	80.85	88.80
50	96.00	97.62	86.02	97.25	98.83
75	99.32	99.43	95.17	99.60	99.85
100	99.90	99.80	98.52	99.93	100
150	100	99.93	99.80	100	100
200	100	99.98	99.98	100	100
250	100	100	100	100	100
300	100	100	100	100	100
400	100	100	100	100	100
500	100	100	100	100	100
750	100	100	100	100	100
1000	100	100	100	100	100

The test statistic g^2 seemed equivalent to K^2, $\sqrt{b_1}$ and W for sample sizes greater than 50. For sample size less than 50, W seemed to be the most powerful. Once again , for the skewed distributions, b_2, had the lowest power up until sample size 100 where it became nearly equivalent in power to the other four statistics. All of the test statistics demonstrated higher power for skewed/peaked distributions than any of the other distribution categories considered earlier. In fact, nearly full power was demonstrated at sample sizes greater than 100 for all test statistics.

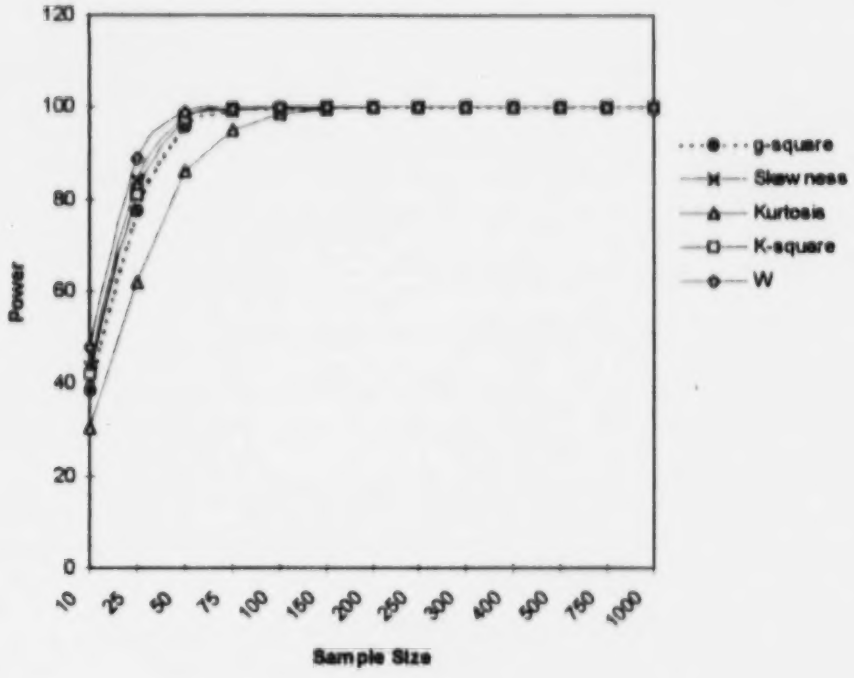

Figure 7: Mean Powers for the Skewed/Peaked Distributions as a Function of Sample Size and Test Statistic

Summary of Power Results

The results showed that the Type I error rate for g^2, $\sqrt{b_1}$, and b_2 were as expected. However, the Type I error rate for K^2 was erratic, with inflated Type I error rates at the .05 and .01 level and lower than expected Type I error rates at the .10 level. The Type I error rate for W was consistently inflated indicating that W may be erroneously detecting departures from normality when the distribution is normal. This inflated power is best seen when the power of W for near normal distributions is considered. For large sample sizes, W had absolute power, yet for large sample sizes these distributions approximate the normal distribution. Therefore, low power in these cases is acceptable. The two moment tests, $\sqrt{b_1}$, and b_2, while having high power for some distributions, do not provide an omnibus test for normality. That is, they only do well when the departure is due to skewness or kurtosis, respectively.

The power estimates showed that, overall, g^2 was sensitive to departures from normality for leptokurtic (peaked) distributions. In fact, its power estimates for these distributions equaled or surpassed the estimates of the competing tests. However, g^2 did not perform as well for platykurtic (flat) distributions or distributions close to normal. For large sample sizes, g^2 also performed well for all distributions except for those that were close in shape to the normal distribution.

Conclusion

The main advantage or strength of g^2 is its conceptual and computational simplicity. Given the formulas and spreadsheet package, the list of critical values could easily be computed. Then, using the output of most any statistical package, g^2 can be computed by hand. Currently, the g^2 value is most easily computed using the SPSS output as SPSS provides both the Fisher estimates for skewness and kurtosis and their respective standard errors. Also, with only minor changes to the reported skewness and kurtosis estimates and with the manual calculation of the standard errors if not provided, g^2 can also be computed using the output of the other major statistical packages. The computed value would then be compared to a table of critical values to determine whether the null hypothesis of normality should be rejected. The relative ease of computing g^2 allows for its use as a supplement to other formal tests of significance which are provided by statistical packages. For example, knowing that W may have inflated Type I error rates, researchers may still request the provided statistic when using SAS or SPSS, but they can also easily compute g^2 as another measure.

Also, the derivation of g^2 does not depend on any distributional assumptions as does K^2. That is, since g^2 is not based on an existing distribution theory, it is empirically derived making it free from restrictive assumptions specific to particular distributions. In addition, as found in the power study, g^2 is sensitive to a wide range of alternative distributions, particularly peaked distributions, having absolute power for many distributions with large n. Therefore, g^2 would be valuable in testing for univariate normality in statistical routines, such as structural equation modeling, where sample sizes are large and where large kurtosis values are problematic. Furthermore, g^2 is unbiased and protects against Type I errors making it the preferred test for normality if the chance of committing a Type I error is the main concern.

The test statistic g^2, however, also has some weaknesses. One of its main disadvantages is its low power with small sample size except for peaked distributions. In addition, for flat distributions, g^2 is not as powerful as the other competing tests, regardless of sample size. Also g^2 appears to be more sensitive to departures due to kurtosis rather than skewness. Furthermore, while it is easy to compute by hand, the determination of the significance of g^2 is currently dependent upon a table of critical values. Therefore, if researchers wish to use g^2, they must have a table of critical values

available.

As with other formal tests of significance, g^2 is not meant to be a replacement to the qualitative information obtained by these graphical representations. While g^2 can tell you that you have a departure from normality, it can not tell you that this departure is due to a single outlier. Therefore, it is recommended that when testing for departures from normality, g^2 should be used as a supplemental quantitative measure of normality to the information obtained by histograms, box plots, stem and leaf diagrams, and normality plots.

Ware and Ferron (1995) noted that g^2 "holds promise as an easily computed and readily available measure of normality." At the onset of this study, it was hoped that g^2 would provide the gold standard for testing for departures for normality. Unfortunately, g^2 did lack power in detecting departures from normality for flat distributions. While traditionally it is peaked distributions that cause problems when the normality assumption is violated, we can not overlook the lack of power this new test showed for non-peaked distributions.

However, regardless of the performance of g^2, the findings in this study cast a doubt on the validity of using W and K^2 as measures of detecting normality. More importantly, the results of this study indicate that g^2 has merit as a test for normality as it is unbiased, easy to compute, readily available, powerful for large sample sizes, and powerful for peaked distributions.

References

Bollen, K. A. (1989). Structural Equations with Latent Variables. New York: John Wiley & Sons, Inc.

Box, G. E. P. (1953). Non-normality and tests for variance. Biometrika, 40, 318-335.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical Statistics and Psychology, 31, 144-152.

Breckler, S. J. (1990). Application of covariance structure modeling in psychology: Cause for concern? Psychological Bulletin, 101, 343-362.

Chou, C. & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. Hoyle (Ed.), Structural Equation Modeling: Concepts, Issues, and Applications (pp. 37-55). Beverly Hills, CA: Sage Publication, Inc.

D'Agostino, R.B., Belanger, A., & D'Agostino, Jr., R.B. (1990). A suggestion for using powerful and informative tests of normality. The American Statistician, 44, 316-321.

D'Agostino, R. & Pearson, E.S. (1973). Test for departure from normality: Empirical results for the distributions of b_2 and $\sqrt{b_1}$. Biometrika, 60, 613-622.

Fisher, R. A. (1973). Statistical Methods for Research Workers. (14th ed.). New York: Hafner Publishing.

Geary, R. C. (1947). Testing for normality. Biometrika, 34, 209-242.

Glass, G. V. & Hopkins, K. D. (1984). Statistical Methods in Education and Psychology. (2nd ed). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage Publications, Inc.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Greenhouse, J. B. & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & Hedges, L. V. (Eds.), The Handbook of Research Synthesis (pp. 383-410). New York: Russell Sage Foundation.

Hayduk, L. A. (1987). Structural Equation Modeling with LISREL. Baltimore: John Hopkins University Press.

Hedges, L. V. & Olkin, I. (1985). Statistical Methods for Meta-Analysis. Orlando, FL: Academic Press.

Hopkins, K.D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. Educational and Psychological Measurement, 50, 717-729.

MacGillivray, H.L, & Balanda, K.P. (1988). The relationship between skewness and kurtosis. Australian Journal of Statistics, 30, 319-337.

Pearson, K. (1895). Contributions to the mathematical theory of evolution, II: Skew variation in homogeneous material. Philosophical Transactions of the Royal Society of London, 91, 343-414.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research: Explanation and Prediction. Fort Worth, Texas: Harcourt Brace College Publishers.

Royston, J. P. (1982). An extension to Shapiro and Wilk's W test for normality to large samples. Applied Statistician, 31, 115-124.

Royston, P. (1992). Approximating the Shapiro-Wilk W test for non-normality. Statistics and Computing, 2, 117-119.

SAS Institute, Inc. (1995). SAS, Release 6.11 [Computer Program]. Cary, NC: SAS Institute, Inc.

Shapiro, S. S> & Wilk, M. B. (1965). An analysis of variance test of normality (complete samples). Biometrika, 52, 591-611.

Shavelson, R. J. (1988). Statistical Reasoning for the Behavioral Sciences. (2nd Ed.). Boston: Allyn and Bacon, Inc.

Tabachnick, B.G. and Fidell, L.S. (1996). Using Multivariate Statistics. (3rd Ed.). New York: Harper and Row, Publishers.

Ware, W. B. and Ferron, J. M. (1995). Skewness, Kurtosis, and Departures from Normality. Paper presented at the Annual Meeting of the North Carolina Association for Research in Education.

West, S. G., Finch, J. G., & Curran, P. J. (1995). Structural equation models with nonnormal variables. In R. Hoyle (Ed.), Structural Equation Modeling: Concepts, Issues, and Applications (pp. 56-75). Beverly Hills, CA: Sage Publication, Inc.

Wolf, F. M. (1986). Meta-Analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA: Sage Publicaitons, Inc.